DOCUMENT RESUME

ED 189 105 TM 800 255

AUTHOR Douglass, James B.
TITLE Applying Latent Trait Theory to a Classroom Examination System: Model Comparison and Selection.
PUB DATE Apr 80
NOTE 56p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Ability: *Achievement Tests: Course Content: Difficulty Level: *Equated Scores: Higher Education: *Latent Trait Theory: *Mathematical Models: Multiple Choice Tests: Test Items
IDENTIFIERS *Rasch Model

ABSTRACT

The three-, two- and one-parameter (Rasch) logistic item characteristic curve models are compared for use in a large multi-section college course. Only the three-parameter model produced clearly unacceptable parameter estimates for 100 item tests with examinee samples ranging from 594 to 1082. The Rasch and two-parameter models were compared for consistency and validity of item calibration and for test equating with examinee samples of 200, 600, 800 and 1082. The Rasch model provided the most consistent and valid equating. For the two-parameter model, a calibration method using average item discrimination ratios provided better results than were obtained by calibrations based on item difficulty regressions. Practical guidelines for others interested in comparing latent trait models for their own applications are provided. Although the focus of the paper is on classroom testing, the methods and recommendations are applicable to testing systems in general. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189105

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

APPLYING LATENT TRAIT THEORY TO A CLASSROOM EXAMINATION SYSTEM: MODEL COMPARISON AND SELECTION

JAMES B. DOUGLASS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Douglass
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Learning and Evaluation Service
Michigan State University

TM 8002255

Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 7-11, 1980. This paper is based on a study supported in part by the Educational Development Program, Michigan State University. The author wishes to thank Cassandra Book, Eric Eisenberg and LeRoy Olson for their varied and valuable help.

ABSTRACT

The three-, two- and one-parameter (Rasch) logistic item characteristic curve models are compared for use in a large multi-section college course. Only the three-parameter model produced clearly unacceptable parameter estimates for 100 item tests with examinee samples ranging from 594 to 1082. The Rasch and two-parameter models were compared for consistency and validity of item calibration and for test equating with examinee samples of 200, 600, 800 and 1082. The Rasch model provided the most consistent and valid equating. For the two-parameter model, a calibration method using average item discrimination ratios provided better results than were obtained by calibrations based on item difficulty regressions. Practical guidelines for others interested in comparing latent trait models for their own applications are provided. Although the focus of the paper is on classroom testing, the methods and recommendations are applicable to testing systems in general.

Item characteristic curve (ICC) models, also called latent trait or item response models, have drawn increasing interest in recent years from both psychometric researchers and practitioners (see Hambleton and Cook, 1977; Wright, 1977; and Hambleton, et al, 1978). The primary appeal of ICC models lies in their potential to solve the problem of deriving comparable scores from different tests given to different people. For example, in Communication 100, an introductory course at Michigan State University, different instructors give different mid-term examinations to different sections. If one section receives lower scores than another, does it mean that the students in that section have less knowledge or a harder test? It would be highly desirable to provide a common scale to which scores from each mid-term could be referenced. Although a common final examination is administered each term, the finals differ across terms. Again a common reference scale would be useful. Students' grades should not depend upon the difficulty of the particular tests they take. Nor should their grades depend on the grades of other students in their particular section. If the potential of ICC models to provide item statistics that are independent of particular examinees and ability scores that are independent of particular items can be realized in the classroom, then score comparability can be achieved in a classroom testing system.

The focus of this paper is on applying ICC models to classroom achievement testing. Three primary objectives of this paper are:

1. To compare the one-, two-, and three-parameter logistic item characteristic curve models for item calibration and test equating in a large multi-section college course.

2. To illustrate a general method for model comparison and selection.
3. To provide practical guidelines for those considering using logistic item response models in their own testing system.

Relatively little ICC research has focused on achievement testing and even less has been addressed to classroom examination. The present study is one of the first to systematically apply ICC models to a classroom examination system. As Bejar, Weiss and Kingsbury (1977) point out, the problem of applying item characteristic curve models to classroom achievement tests is significantly different from the problem of applying the models to aptitude tests. ICC models assume that a single trait will account for all systematic differences in examinees' scores. As Bejar, Weiss and Kingsbury correctly observe, aptitude test item pools can typically be constructed and refined to meet the unidimensionality assumption of ICC models quite well. However, a classroom achievement test must fairly represent the entire course content. There is much less freedom to delete areas of content in classroom tests than there is in aptitude tests.

Fortunately the question of interest is not whether the assumptions of ICC models are met or not. In the real world, assumptions are never met. The practical questions are 1) how well are the assumptions met and 2) what effect does this have on the desired properties of the model? Robustness of the models to violations of their assumptions is the central concern.

There is a problem resulting from the robustness issue which has been largely ignored in the ICC literature. As an illustration of the problem consider Wright's (1977, p. 113) procedure for designing a "best test."

His procedure involves only psychometric properties of the items in the test and guesses about the mean and dispersion of the abilities of persons to be tested. There is no mention of item content. Of course, if an item pool is truely unidimensional then the content of any particular item may be irrelevant. But a model may give consistent and valid results either because its assumptions are met or because it is robust with respect to violations of its assumptions. A model may show evidence of fit despite a lack of unidimensionality. Therefore the content validity of a test is still a relevant concern in test development. The use of latent trait models with classroom achievement tests does not relieve the test constructor of the task of developing a test with balanced representation of course content.

Lord and Novick (1968, p. 383) suggest the following four steps for investigating the practical utility of a model:

1. Estimate the parameters of the model, assuming it is true (see Section 17.9).
2. Predict various observable results from the model, using the estimated parameters.
3. Consider whether the discrepancies between predicted results and actual results are small enough for the model to be useful ("effectively valid") for whatever practical application the investigator has in mind.
4. If in step 3 the discrepancies were considered too large, then it may be useful to compare them with the discrepancies to be expected from sampling fluctuations.

The research strategy in this study is based on a paper by Douglass (1979), in which he elaborates and extends the Lord and Novick procedure to include comparisons of different models for consistency, validity, and utility. The basic steps Douglass suggests are:

1. Identify the models to be compared.

2. Obtain different sets parameter estimates for each model, assuming each is a true representation of reality. These estimates should be obtained in the actual situation where the selected model will be used and should reflect any sources of error that will be operating in the practical setting. In addition to naturally occuring variation, variation can be introduced artificially to simulate the practical situation. For example, different size samples can be investigated by splitting a large sample into sub-samples. This will result in discrepant estimates for each model.
3. Observe the effects of the discrepant estimates for each model on each practical situation of interest.
4. If any given model does not give sufficiently consistent results, then possible sources of error should be investigated. Sources of estimate instability include sampling error or bias, estimation procedure error and violation of model assumptions.
5. Compare the results of the models which give consistent results or can be made to give consistent results to one another.
6. Investigate the patterns of model convergence and divergence along with any available criteria to aid in model selection.

Method

Sample

The data is taken from an introductory communication course offered at Michigan State University. Most students taking the course are freshmen, some are sophomores and a few are juniors or seniors. The course is divided into sections of fifty to seventy students taught by different instructors. Each instructor constructs his/her own mid-term examination. All sections are given a common final examination. The final examination is constructed by the course coordinator from an item pool of approximately 600 items. Each final examination consists of 100 four-option multiple choice items. Test construction is based on principles derived from classical test theory. Items are preferred with high upper-lower discriminations and with difficulties of approximately .30. Each final examination reflects a balanced coverage of the ten content areas of the course. For a more detailed description of the course, the item pool, the test construction, as well as administrative and instructor training concerns see Eisenberg and Book (1980).

Data for this study comes from the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 final examinations taken by 947, 820, 594 and 1082 examinees, respectively. For further analysis half samples and independent random samples of 200, 600 and 800 examinees were selected from the Fall 1979 sample. Each examination consisted of 100 items. The Spring 1979 examination was chosen to provide the base scale for calibrating abilities and item difficulties. Calibration required that the Spring 1979 examination share items with each of the other three examinations. Between the Spring 1979 examination and the Fall 1978, Winter 1979 and Fall 1979 examinations there were 49, 53 and 43 items in common, respectively.

Models

The three models selected for comparison were the three-, two- and one-parameter or Rasch, logistic item response models. Each of these models specifies a particular mathematical relationship between person ability and the probability of a correct response to a given item (see Hambleton and Cook, 1977 for a good introduction to ICC models). The three-parameter model can be written:

$$P_g(\theta, a_g, b_g, c_g) = c_g + (1 - c_g)\left(1 + e^{-1.7\, a_g (\theta - b_g)}\right)^{-1}, \quad (1)$$

where $P_g(\theta, a_g, b_g, c_g)$ is the probability of a correct response to item g by a person of ability θ,

c_g is the probability of examinees of extremely low ability giving a correct response to item g, i.e. the lower asymptote of the ICC,

b_g is the point on the ability scale where the slope of the ICC is maximum and is usually called the item's difficulty,

a_g is proportional to the slope of P_g at $\theta = b_g$ and is called the item's discrimination,

and e is the base of the natural logarithms and is approximately equal to 2.718.

The two-parameter model is identical to the three-parameter model when all of the lower asymptotes (c_g) are fixed at a constant, usually zero. The two-parameter model when $c_g = 0$ can be written:

$$P_g(\theta, a_g, b_g) = (1 + e^{-1.7\, a_g (\theta - b_g)})^{-1}, \qquad (2)$$

where the notation is defined above.

If the model is further simplified by substituting $\bar{a}$, the mean of the item discriminations in a test, for a_g, the one-parameter model is obtained:

$$P_g(\theta, b_g) = (1 + e^{-1.7\, \bar{a} (\theta - b_g)})^{-1}. \qquad (3)$$

A version of the one-parameter logistic model was developed independently by Rasch (1960). This version, called the Rasch model, is usually written with the constant 1.7 $\bar{a}$ incorporated into the θ (b) scale and can be written:

$$P_g(\theta', b_{\acute{g}}) = (1 + e^{-(\theta' - b_{\acute{g}})})^{-1}. \qquad (4)$$

Estimation

The computer program LOGIST (Wood, Wingersky and Lord, 1976) was used to estimate ability and item parameters for the two- and three-parameter logistic models. Although LOGIST can be used to estimate parameters for the one-parameter model, BICAL (Wright, Mead, and Bell, 1979) was used. One trivial difference between the programs is that the estimates for LOGIST are based on equations (1) - (3) while BICAL is based on equation (4). More importantly the estimation procedure used in BICAL was developed specifically for the Rasch model and should provide better estimates for that model than LOGIST does. LOGIST is primarily designed for researchers while BICAL is more user oriented and therefore has many convenient and useful features not found in LOGIST.

Scaling

The degree of uniqueness of the person ability and item parameter scales is determined by the item response models. The critical issue is the extent to which the parameters of the model are determined. For the three-parameter model, consider a single item, g, where P_{g1} (θ_1, a_1, b_1, c_1,) represents one scaling and P_{g2} (θ_2, a_2, b_2, c_2,) represents another scaling. We will consider what $P_{g1} = P_{g2}$ implies about the relationship of the two scales.

Notice immediately from equation (1) that the minima of the P_g's occur as their respective θ's approach negative infinity. The minimum of P_{g1} is C_1, and the minimum of P_{g2} is C_2, but we are considering the same item so $C_1 = C_2$. In other words the lower asymptotes of the item characteristic curves are completely determined.

Now consider the case where P_{g1} and P_{g2} equal a constant greater than C_g. For the probabilities to be equal, it can be seen from (1) and the fact that $C_1 = C_2$ that:

$$a_1 (\theta_1 - b_1) = a_2 (\theta_2 - b_2). \tag{5}$$

Rearranging terms algebraically we find that

$$\theta_1 - \frac{a_2}{a_1} \theta_2 = b_1 - \frac{a_2}{a_1} b_2. \tag{6}$$

But we are considering a single item with $P_{g1} = P_{g2}$ being held constant so that θ_1, θ_2, a_1, a_2, b_1 and b_2 are all constant. Therefore there exists a

constant K equal to each side of equation (6). So we can write

$$\theta_1 = K + \frac{a_2}{a_1} \quad \theta_2 \tag{7}$$

and

$$b_1 = K + \frac{a_2}{a_1} \quad b_2. \tag{8}$$

Notice in the Rasch Model that (7) and (8) reduce to

$$\theta_1' = K + \theta_2' \tag{9}$$

and

$$b_1' = K + b_2' \,, \tag{10}$$

since the a's are not modeled.

The LOGIST program solves the problem of the indeterminacy in the person ability and item difficulty scales by setting the average person ability to zero and the standard deviation of ability to approximately one on any given computer run. Therefore, two separate LOGIST runs with different groups of examinees will result in scales that are linear transformations of one another. In this study we have different examinees but common items across tests so equation (8) is the appropriate one to examine, where now b_1 is on the scale defined on one test and b_2 is on the scale defined by another test.

Calibration requires that we estimate the transformation represented by (8). This can be done in several ways. Perhaps the most obvious way is to use the item difficulty estimates to estimate the parameters of the regression equation.

$$b_1 = \alpha + \beta\, b_2, \tag{11}$$

where we have as many sets of $(\hat{b}_1, \hat{b}_2)$ as we have items in common between the two tests.

A alternative method of estimating α and β, developed by the author, uses the item discriminations and difficulties. This method was obtained by comparing equations (8) and (11), and noticing that they are identical if $K = \alpha$ and

$$\frac{a_2}{a_1} = \beta\ . \tag{12}$$

Then a reasonable estimate of β is the average of the ratios $\frac{\hat{a}_2}{\hat{a}_1}$. Here we have as many $(\hat{a}_1, \hat{a}_2)$ as we have items in common between the two tests. An estimate of α can be obtained by substituting $\hat{\beta}$ for β, the mean of the b_1's for b_1 and the mean of the b_2's for b_2 in equation (11) and solving for α. Each method will result in a generally different linear equation for converting abilities and difficulties from the second scale to the first. Other methods could be developed that also use both the difficulties and discriminations simultaneously. The best method of combining the information contained in the discriminations and difficulties is not obvious.

The BICAL computer program solves the indeterminacy of the Rasch scale by setting the mean of the items in any given test to zero. By substituting the mean of the b_1''s for b_1' and the mean of the b_2''s for b_2' and solving for K in equation (10) we get

$$K = \bar{b_1'} - \bar{b_2'} . \tag{13}$$

Equating

Once scaling has been accomplished, test equating becomes simple. If ability scores are being compared then the same transformation that was applied to item difficulties can be applied to abilities to form ability measures on the same scale.

If raw score equating is desired then a further complication is introduced; it is necessary to set up a correspondence between raw scores in different tests through the ability estimates. One method of equating uses true score equating applied to raw scores. The true score on a test for a given ability is simply the expected value of the observed score at that ability level. This is simply the sum of the item characteristic curves that the ability, i.e.,

$$T(\theta) = E(X|\theta) = \Sigma P_g(\theta) , \tag{14}$$

where the summation is over all the items in a test. Equating is simply a matter of tabling or graphing the relation of $\Sigma \hat{P}_g(\theta)$ on one test to that of $\Sigma \hat{P}_g(\theta)$ on a second test for many fixed values of θ, where

$\hat{P}_g(\theta) = P_g(\theta, \hat{a}_g, \hat{b}_g, \hat{c}_g)$ (see Lord, 1977).

Procedure

Separate estimates for each of the samples mentioned above were obtained for the Rasch model using BICAL. Examinees with less than 40 of 100 items correct or with person t fits above 2.00 (see Wright, Mead, and Bell, 1979, p. 15) were deleted from the item calibration sample. From 1/2% to 5% of the examinees in a given sample were removed under these criteria. It was felt that persons with scores of less than 40 were likely to be guessing on many items and thereby adding error to item calibrations. Persons with large t fit values were removed since they were demonstrating unusual response patterns which could adversely affect item parameter estimation.

All item difficulties except those for the half samples of Fall 1979 were transformed to the scale defined by the Spring 1979 examination using equations (10) and (13). The person abilities were then transformed to a standard scale designed so that the mean ability of the Spring 1979 examinees is 500 and the standard deviation is 100. The transformation equation is

$$S = \frac{\theta - 1.04}{.60}(100) + 500 , \qquad (15)$$

where S is the scaled score,

θ is estimated ability on the Spring 1979 scale,

1.04 is the person ability mean on the Spring 1979 scale,

and .60 is the ability standard deviation on the Spring 1979 scale.

Separate estimates for each sample were obtained for the two-parameter model using LOGIST. Person abilities and item difficulties and discriminations for all samples except the Fall 1979 half samples were transformed to the scale determined by the Spring 1979 examination. The regression method using item difficulties and the average ratio method using item discriminations and average difficulties were used to estimate α and β in equation (11). This resulted in two different transformations for the two-paramter model. The additional transformation

$$S = (100)\,\theta + 500 \qquad (16)$$

was applied to the scaled estimates to facilitate comparisons between the Rasch and two-parameter models.

LOGIST was used to obtain separate estimates for the three-parameter model on the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 total samples. No calibrations were done.

The Fall 1979 sample of 1082 examinees was divided into two samples of size 541 in order to make it possible to obtain a means of comparing the validity of the various models for equating. One sample was calibrated to the other using only those 43 items that were used to calibrate the various Fall 1979 samples to the Spring 1979 scale. The calibrated abilities were used in equation (14) for true score to true score equating.

Results

Parameter Estimation

The three-parameter model has unacceptable lower asymptote estimates (C_g). Lord (1975) has shown that C_g is poorly estimated when ($b_g - 2/a_g$) is less than negative two. Consequently when ($b_g - 2/a_g$) ≤ -2 LOGIST fixes C_g at the average of the other unfixed C's. For the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 examinations 98, 98, 98, and 99 of 100 items were fixed at the average of the C's of the remaining 2, 2, 2, and 1 items. The averages were .25, .205, .13 and .075, respectively. In effect the three-parameter model was reduced to an inconsistent two-parameter model. Therefore, the three-parameter model was given no further consideration.

On the whole, estimation of parameters for the two-parameter model was acceptable. Discriminations can approach zero or infinity so LOGIST places the restriction $.01 \leq a_g \leq 2.0$. No more than one of 100 items was fixed at .01 and no items approached 2.0 in any two-parameter LOGIST run. There were few implausable item difficulty estimates in any examination. Most samples had 0, 1 or 2 of 100 items with estimated difficulties more than six standard deviations below the average person ability and only one had as many as four items. Even in the unlikely event that such estimates are accurate, such items are practically useless for measurement in the present situation. No items had implausibly high difficulties. LOGIST applied to the two- and three-parameter models has the disconcerting property of allowing parameter estimates to obtain absurd values. Clearly such items do not fit the model.

Rasch model estimates obtained from BICAL were all plausible. Item tests of fit provided by BICAL are highly dependent on sample size. For example,

the samples of 200, 600, 800 and 1082 from the Fall 1979 examination identified 8, 13, 25 and 29 items respectively as having an overall t fit greater than 2.0. It is clear that the fit statistics are more appropriate for item selection than for testing model fit unless some means of taking sample size into account is found.

In conclusion estimates were successfully obtained for the Rasch and two-parameter models. The next two sections on calibration and equating provide information on the consistency, accuracy and usefulness of the estimates.

Calibration

For the Rasch model the Fall 1978, Winter 1979, Fall 1979 examinations and random samples of 200, 600 and 800 of the Fall 1979 examination were separately calibrated to the Spring 1979 scale by the method previously described. Table 1 presents the calibrating constants, residual and standardized residual statistics and the numbers of items in the calibrating samples.

The residual is simply the difference for the common items between the actual Spring 1979 item difficulties and the item difficulties calibrated from another test scale. The residual mean is necessarily zero. A totally error-free calibration would result in a residual standard deviation of zero.

The standardized residual is simply a residual divided by the standard error of that residual. The standard error of a residual is

$$S_R = (S_1^2 + S_2^2)^{\frac{1}{2}} \qquad (17)$$

Table 1

Rasch Model Calibration

	Academic Term					
	Fall 78	Winter 79	Fall 79	Fall 79	Fall 79	Fall 79
Person Sample	947	820	200	600	800	1082
Calibrating Constant, K	.072	.131	.179	.170	.172	.172
Residual Mean	0	0	0	0	0	0
Residual S.D.	.39	.31	.37	.37	.36	.37
Standardized Residual Mean	.06	.02	-.08	-.08	-.05	-.05
Standardized Residual S.D.	3.01	2.30	1.83	2.57	2.70	2.93
Number of Calibrating Items	49	53	43	43	43	43

The form of the equation which transforms item difficulty estimates from one scale ($\hat{b}$) to the Spring 1979 scale ($\hat{b}_g$) is:

$$\hat{b}_S = \hat{b} + K.$$

where S_1^2 is the error variance for an item in test 1 and S_2^2 is the error variance for the same item in test 2. In general, different items will have different residual standard errors.

The mean of the standardized residuals should be approximately zero and the standard deviation should be about 1.0. It can be seen in Table 1 that the means are all close to zero, with all discrepancies less than .1. However, the standard deviations are up to three times larger than expected. This implies that there is more error in the calibrating links than can be explained by the model.

For the two-parameter model, both methods of calibrating that were previously described were separately applied to the various samples. As with the Rasch model the Spring 1979 examination provided the common scale. Table 2 contains the calibrating constants estimated from the regression of item difficulties (B Method) and from the ratios of item discriminations (A Method). It also contains correlations between the difficulty estimates and numbers of common items between Spring 1979 and each of the other examinations that were used for calibration. Notice that two items in the Fall 1978 examination and one item in the Winter 1979 examination were not used for calibration. Each item's difficulty was at least 25 standard deviations below the person ability mean. All three greatly and adversely affected the calibrating equations, so they were removed and the calibration equations were again calculated.

Table 2

Two-Parameter Model Calibration

	Academic Term					
	Fall 78	Winter 79	Fall 79	Fall 79	Fall 79	Fall 79
Person Sample	947	820	200	600	800	1082
B Method Calibrating $\hat{\alpha}$	-.986	-.173	-1.094	-.689	-.812	-.707
B Method Calibrating $\hat{\beta}$	.991	.942	.297	.579	.485	.563
A Method Calibrating $\hat{\alpha}$	.155	.018	.522	.154	.178	.178
A Method Calibrating $\hat{\beta}$	1.144	1.089	1.295	1.132	1.115	1.138
Correlation of $\hat{b}$'s	.80	.87	.46	.58	.51	.57
Number of Calibrating Items	47	51	43	43	43	43

The forms of the equations to transform item parameter estimates from one scale ($\hat{a}$, $\hat{b}$) to the Spring 1979 scale ($\hat{a}_S$, $\hat{b}_S$) are

$$\hat{b}_S = \hat{\alpha} + \hat{\beta}\,\hat{b}$$

and

$$\hat{a}_S = \hat{a} \,/\, \hat{\beta}.$$

Notice from Table 2 that there is a very great discrepancy between the calibration equations based on regressed difficulties and those based on average discrimination ratios for all Fall 1979 samples. Also notice that the Fall 1979 sample of 200 results in equations that are quite different from the larger samples, particularly for the difficulty method. The effects of these discrepancies on equating will be seen in the next section.

One half of the Fall 1979 examination sample was calibrated to the other half sample. Table 3 contains the Rasch calibrating statistics. Since BICAL sets the item means to zero and both halves are actually the same examination, the size of the equating constant is an indication of error. The standardized residual mean and the standard deviation are close to the expected values of zero and one, respectively.

The two-parameter model calibrating statistics for the calibration of one half of the Fall 1979 examination to the other are in Table 4. The two sets of calibration equation constants are quite different, considering that both tests are the same and only the examinees differ.

Equating

The Fall 1978, Winter 1979, Spring 1979 and random 200, 600, 800 and total examinee samples of the Fall 1979 examination were equated to a standard reference scale. The standard scale was defined to have a mean of 500 and standard deviation of 100 for the Spring 1979 examinee sample. As described previously, linear transformations of the calibrated Rasch and two-parameter model abilities were used to transform the abilities to the reference scale.

Table 3

Rasch Model Calibration for Fall 1979
Half Samples

Calibrating Constant, K	.034
Residual Mean	0
Residual S.D.	.17
Standardized Residual Mean	.01
Standardized Residual S.D.	1.15
Number of Calibrating Items	43

Table 4

Two-Parameter Model Calibration for Fall 1979
Half Samples

B Method Calibrating $\hat{\alpha}$	-.461
B Method Calibrating $\hat{\beta}$	.696
A Method Calibrating $\hat{\alpha}$	-.088
A Method Calibrating $\hat{\beta}$	.933
Correlation of $\hat{b}$'s	.77
Number of Calibrating Items	43

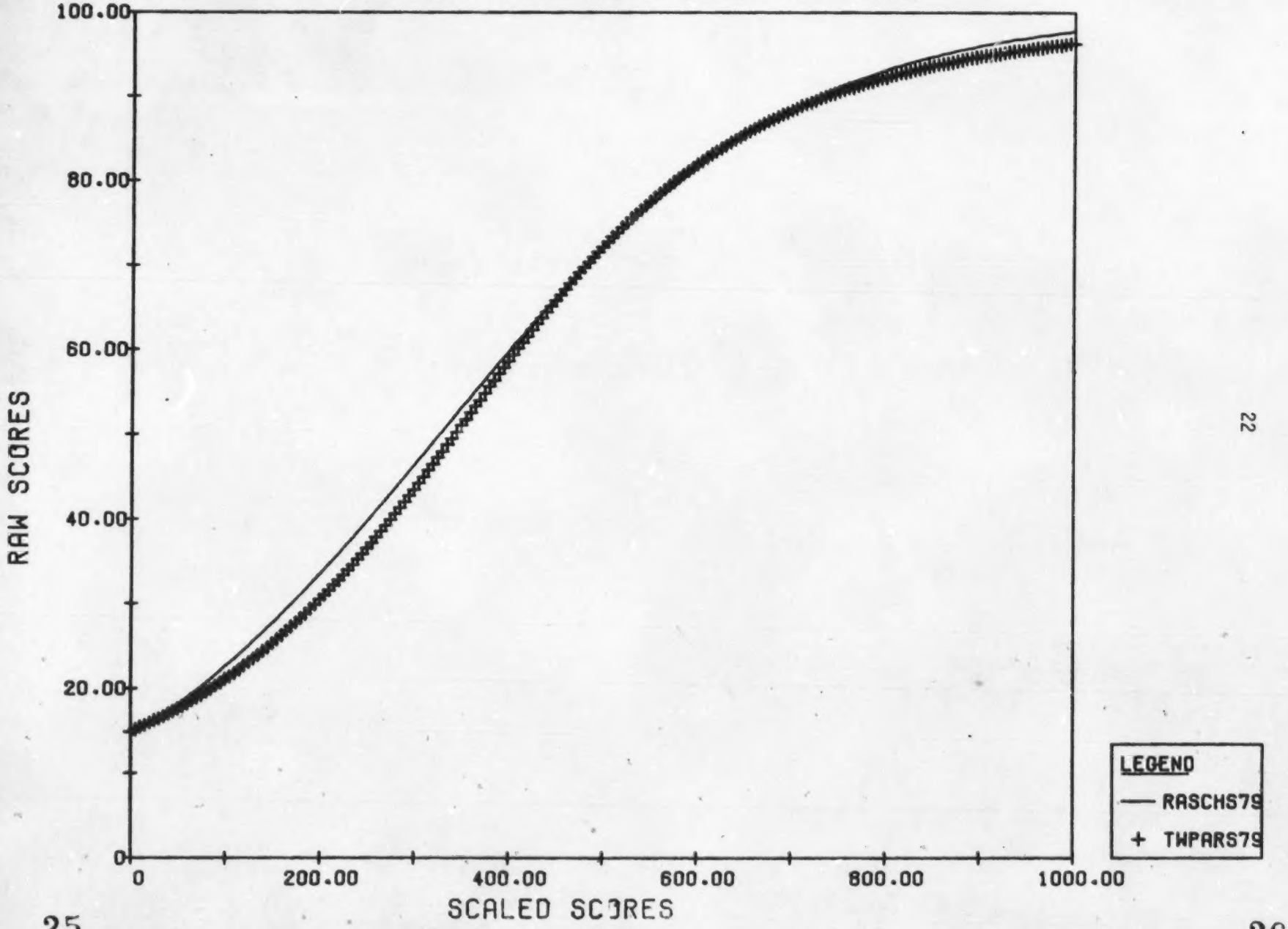
In the figures that follow, axes labeled "scaled scores" correspond to the standard reference scale and the axes labeled "raw scores" correspond to expected raw scores (true scores). Figure 1 provides an indication of the degree of correspondence between the reference scales determined by the Rasch and two-parameter models. It can be seen that the scaling has resulted in reasonably equivalent scales.

Figures 2 through 7 compare the equatings of the Rasch and two scalings of the two-parameter models for the Fall 1978, Winter 1979, Fall 1979 and random 200, 600, and 800 examinee samples of the Fall 1979 examination, respectively.

The variable labeled BMTH in the figures is the equating influenced by the calibration method which uses the regression of item difficulty values. The variable labeled AMTH reflects the alternate calibration method which estimates the regression weights by the average ratio of the item discrimination values. In Figures 2 through 7 the solid line represents the Rasch equating, the plus marks represent the equating using the two-parameter model calibrated by the difficulty regression method, and the circles represent the two-parameter equating with calibration by the alternate method.

An investigation of Figures 2 and 3 reveals that there is very little difference between the three equating methods. Figure 4 is more interesting. The Rasch and two-parameter model with calibration accomplished by regressing item difficulties is strikingly different. For example, the scaled score of 600 corresponds to a raw score of 79.0 for the Rasch Model and raw scores of 93.5 and 79.9 for the two different methods of calibrating the two-parameter

FIGURE 1: SPRING 1979 EQLATING



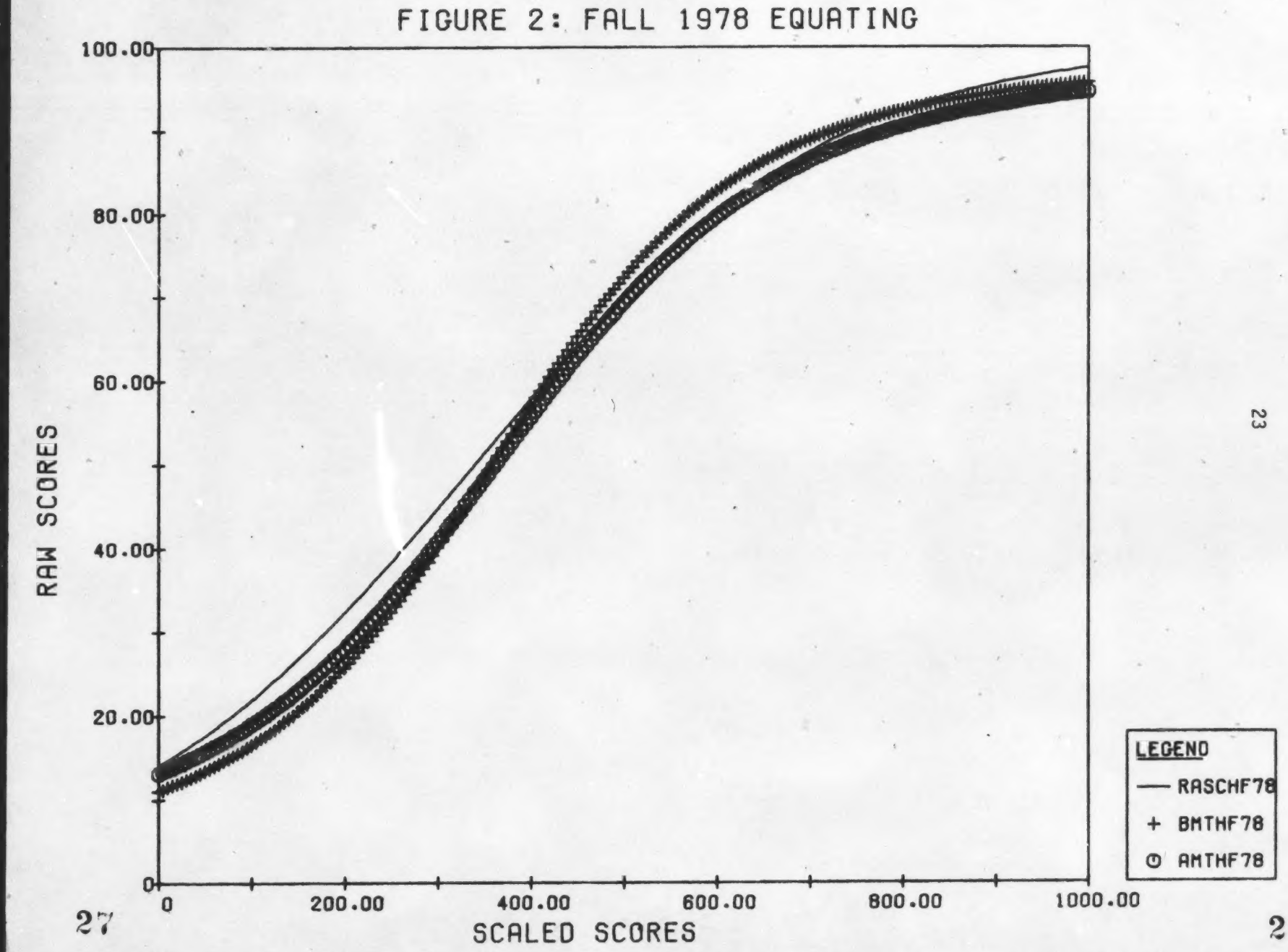


FIGURE 2: FALL 1978 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCHF78
BMTHF78
AMTHF78

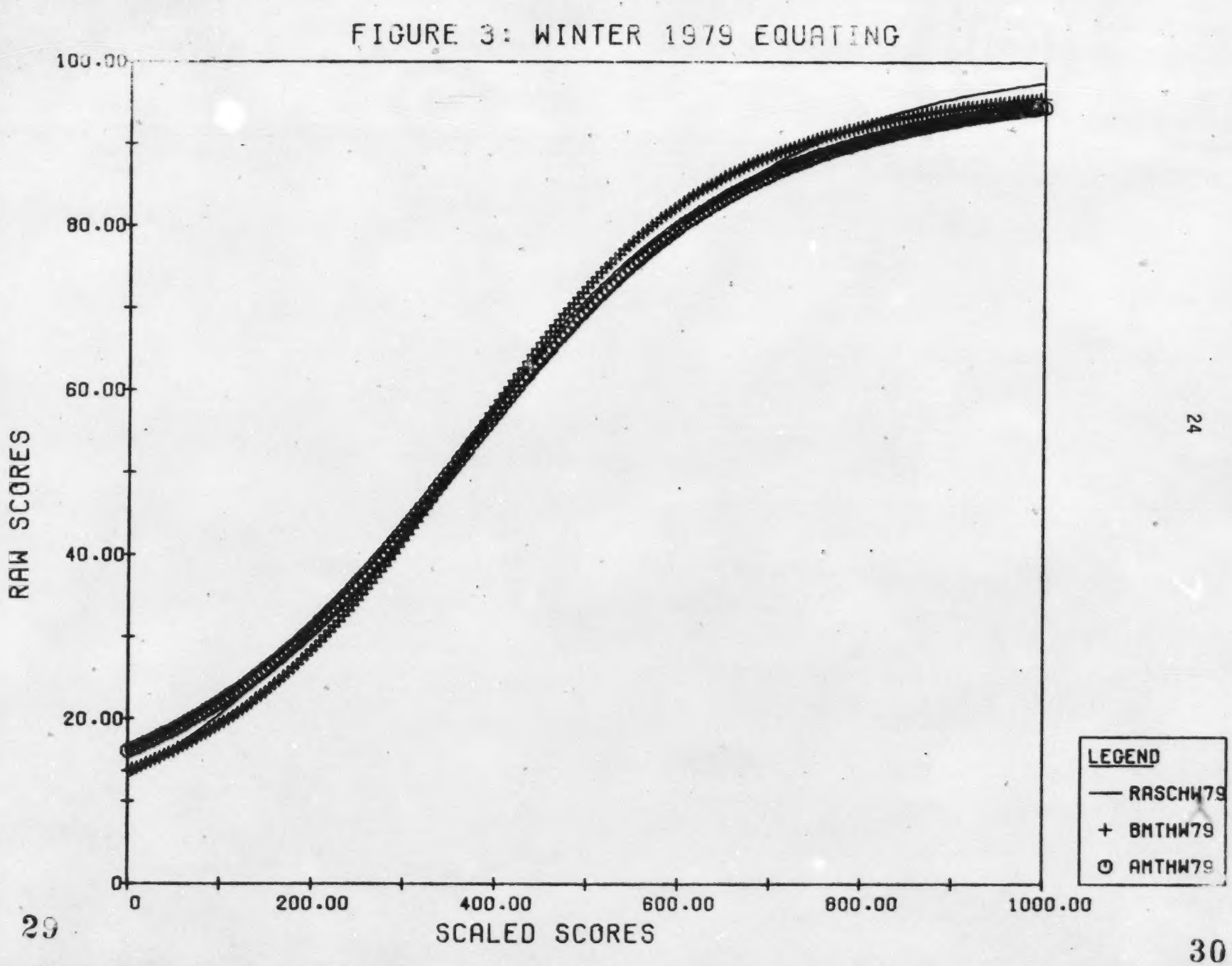


FIGURE 3: WINTER 1979 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCHW79
BMTHW79
AMTHW79

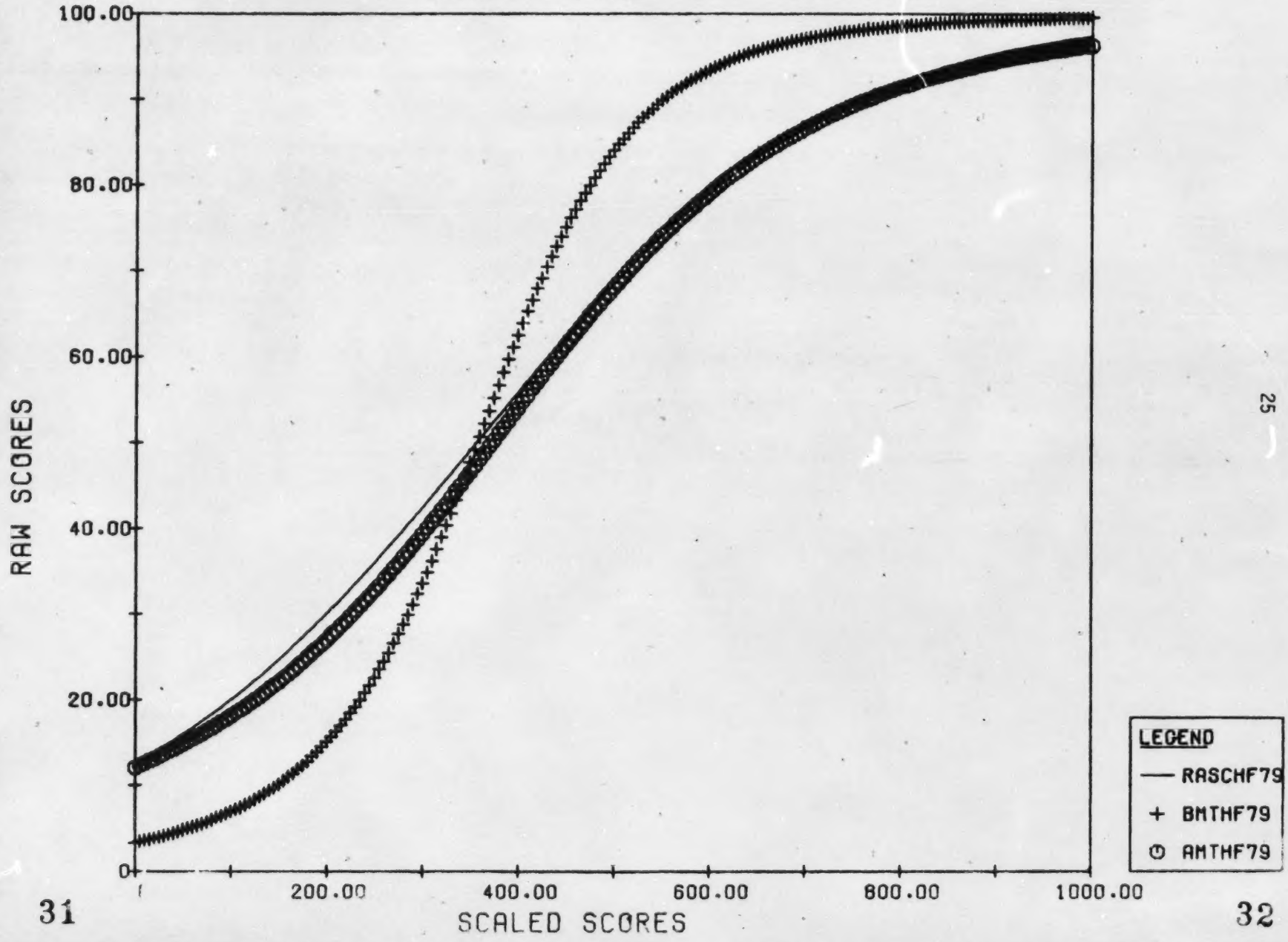


FIGURE 4: FALL 1979 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCHF79
BMTHF79
AMTHF79

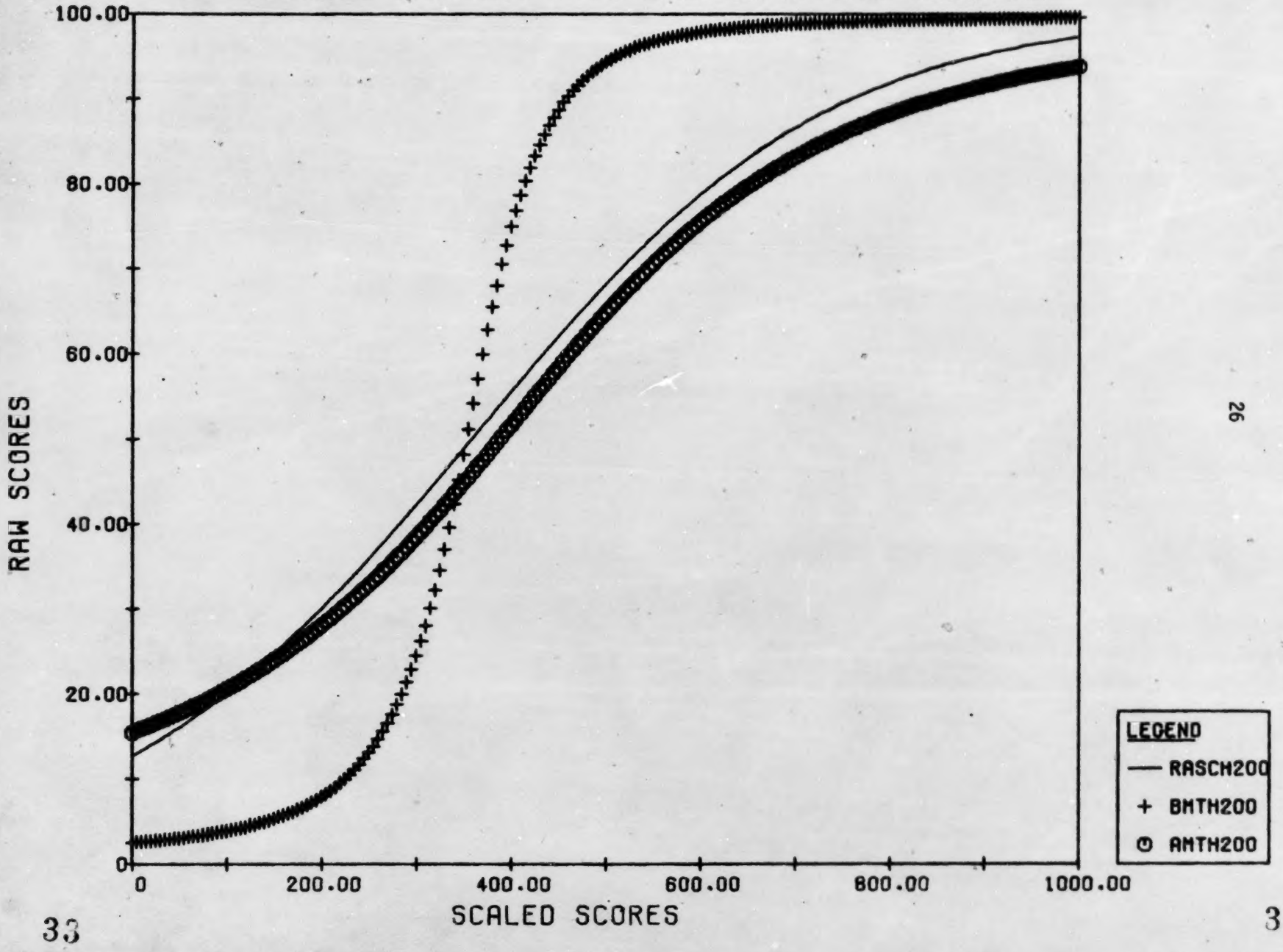


FIGURE 5: FALL 1979 SAMPLE OF 200 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCH200
BMTH200
AMTH200

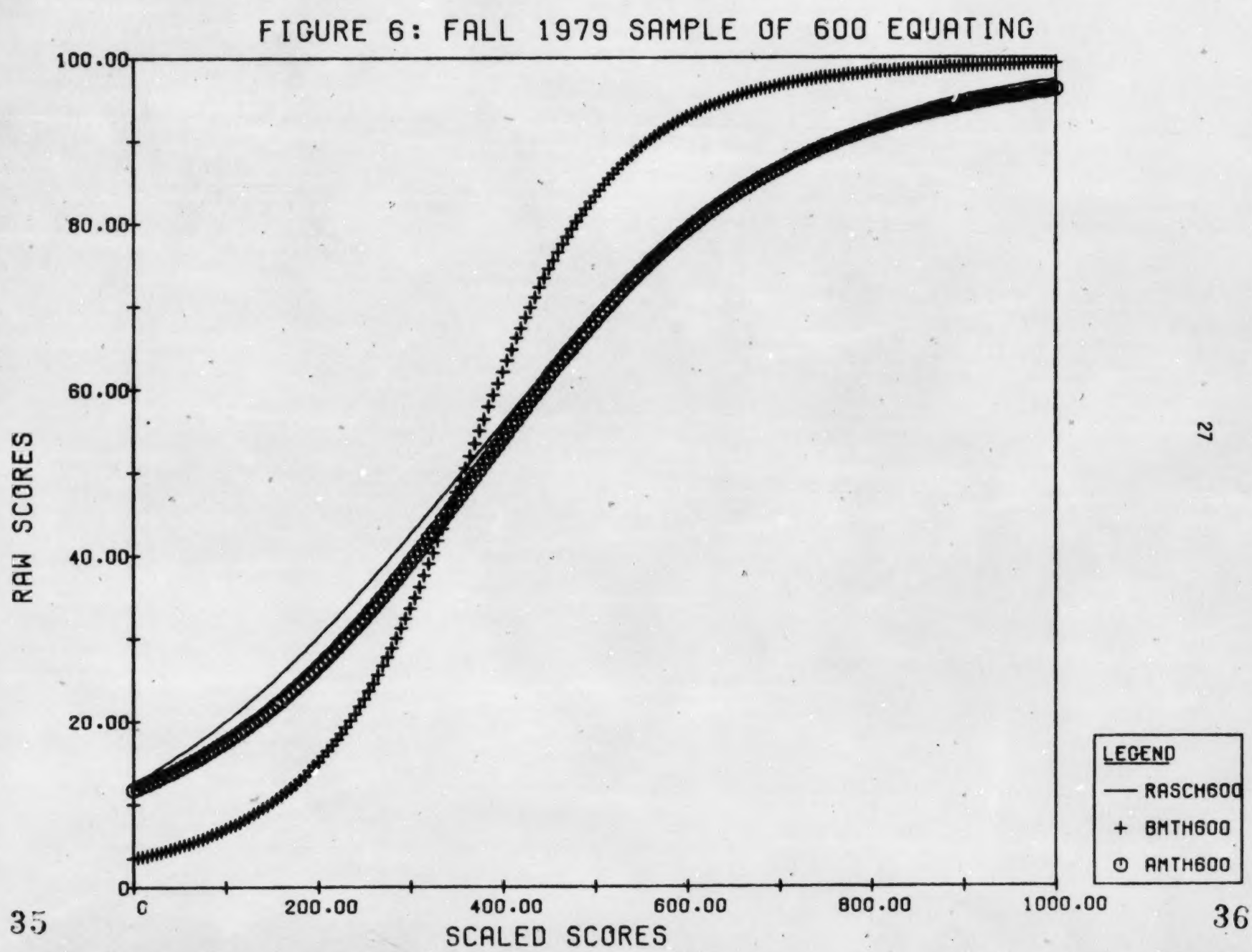


FIGURE 6: FALL 1979 SAMPLE OF 600 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCH600
BMTH600
AMTH600

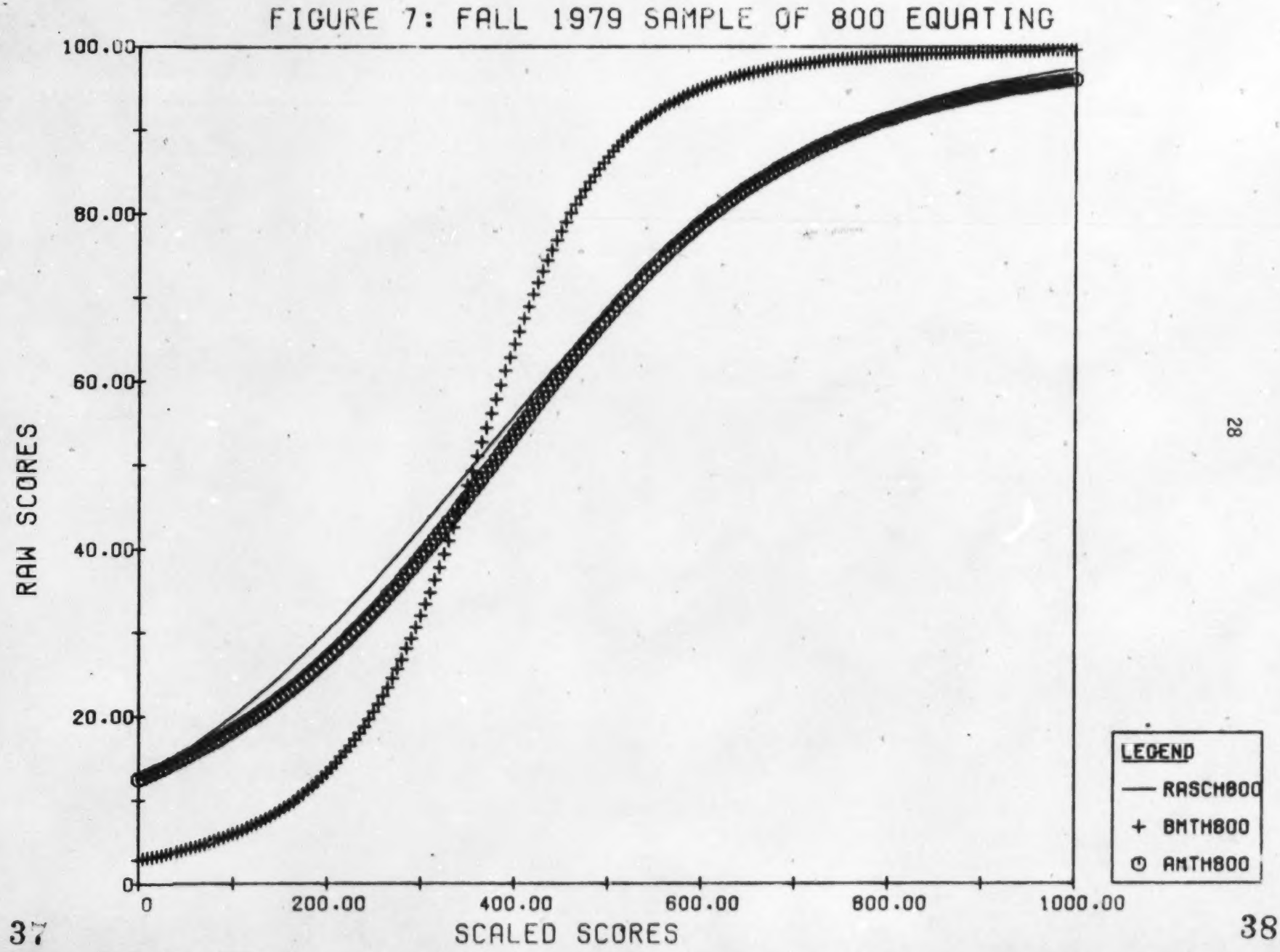


FIGURE 7: FALL 1979 SAMPLE OF 800 EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
RASCH800
BMTH800
AMTH800

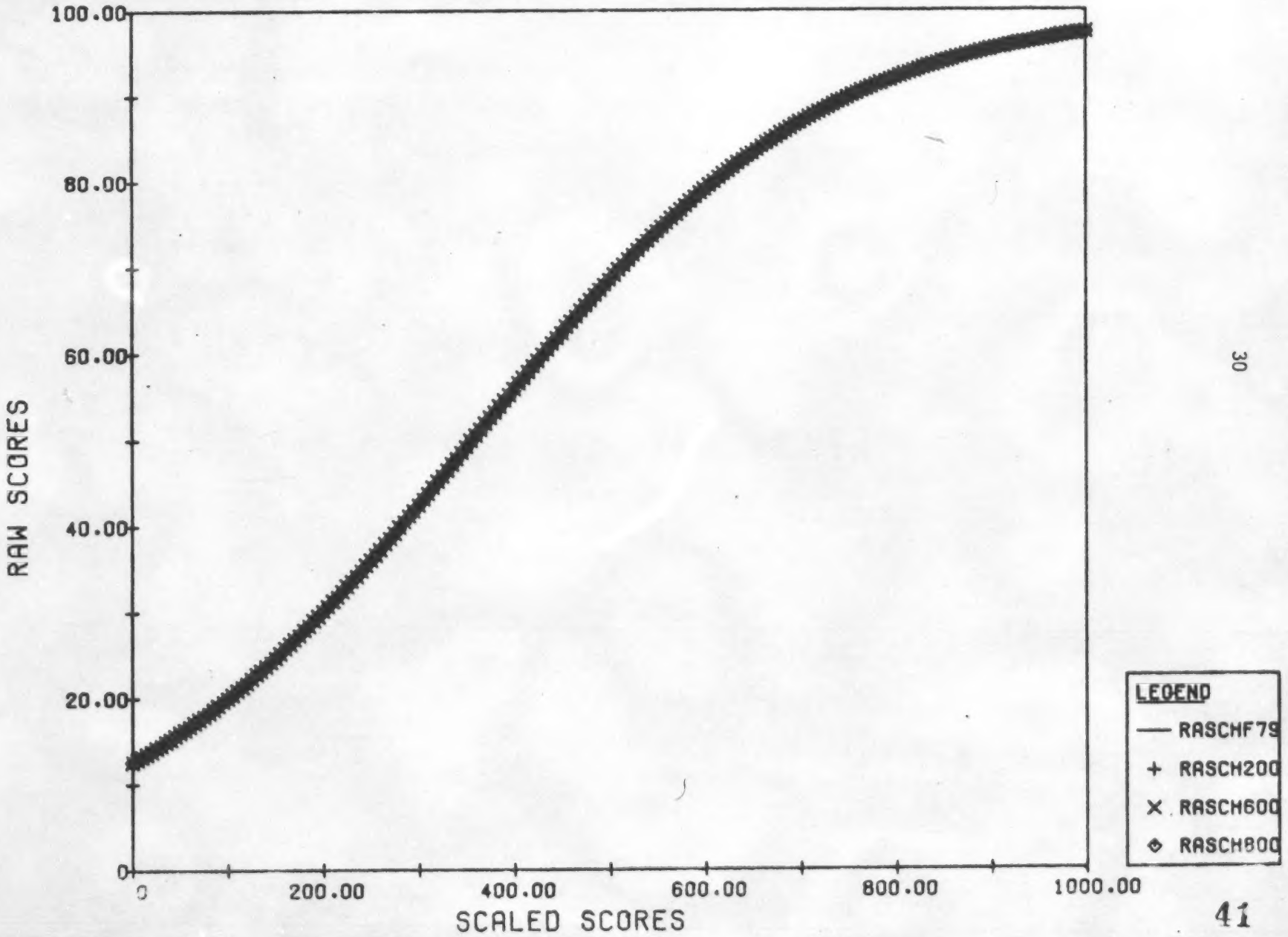
model. Similar patterns are seen in Figures 5, 6, and 7 for samples of 200, 600 and 800 for the Fall 1979 examination. The equating of the sample of 200 examinees has the least consistent results across models.

In Figures 8, 9 and 10 the equating consistency within models is shown for the various size samples of the Fall 1979 examination. The Rasch Model in Figure 8 shows remarkable consistency. The equating lines are virtually indistinguishable. The two-parameter model calibrated by item difficulty regression shows much more divergence in Figure 9, particularly for the sample of 200 examinees. The two-parameter model calibrated by the alternative method displays good consistency in Figure 10. Only the sample of 200 examinees shows any departure from the others.

The inconsistent equating results across models raises the obvious question. Which, if any, of the models are giving the correct equating? One method of investigating this question is to split a sample of examinees in half and equate a test to itself using only a subset of items for the calibrating link. This was done for the Fall 1979 sample. The two samples were calibrated as though they were samples from different tests with only the 43 items used in the original calibration in common. The tests were then equated by the true score method previously described. The equatings should result in a 45^{o} line through the origin.

In Figures 11 and 12 the solid line is the 45^{o} criterion line. In Figure 11, the two-parameter model equatings are shown. It can be seen that both diverge from the proper equating with the alternative method for calibrating consistently giving the better results. In Figure 12 it can be seen that the Rasch model equating and the criterion line are

FIGURE 8: RASCH EQUATING FALL 1979 SAMPLES



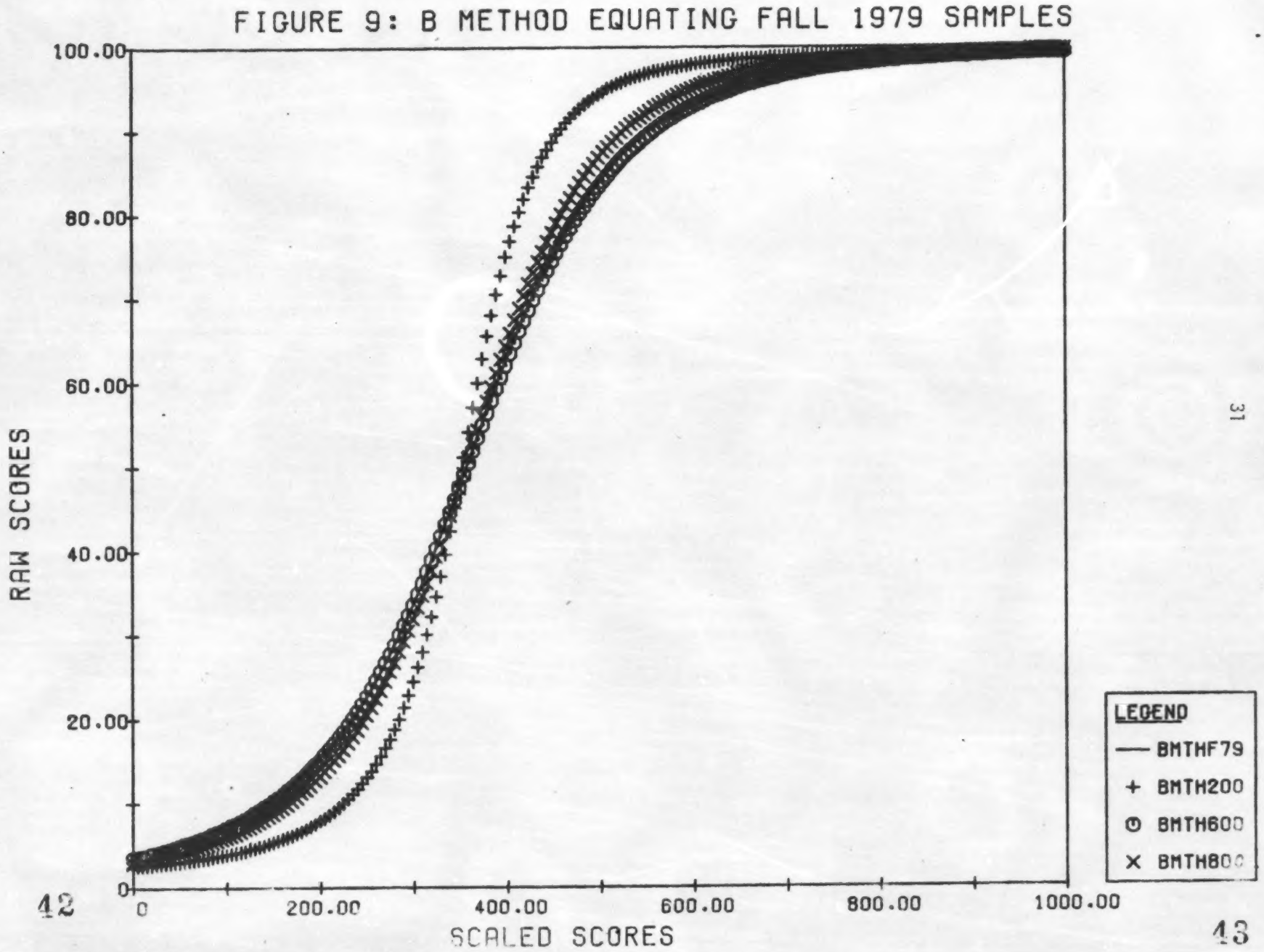
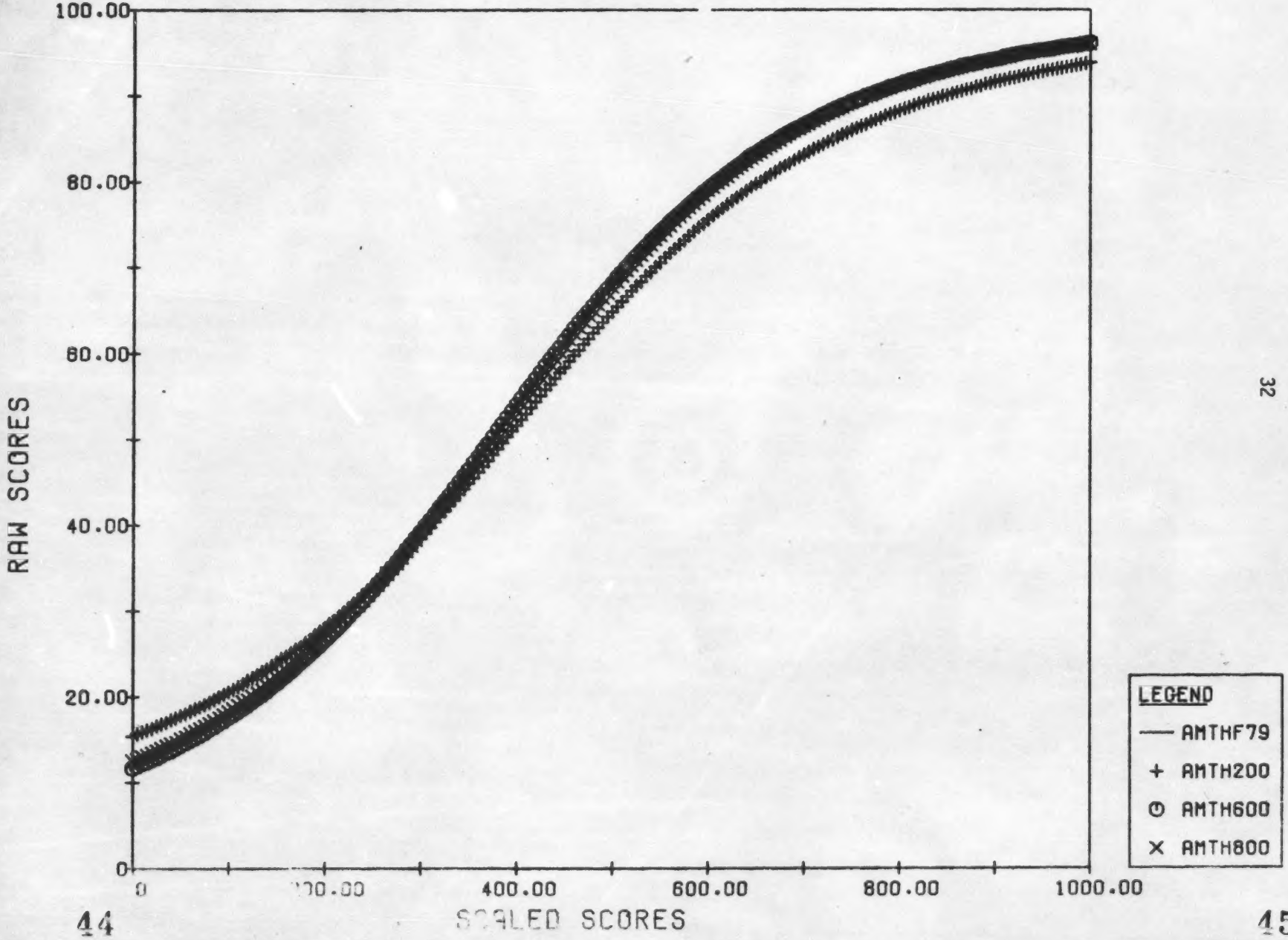


FIGURE 9: B METHOD EQUATING FALL 1979 SAMPLES
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
200.00
400.00
600.00
800.00
1000.00
SCALED SCORES
LEGEND
BMTHF79
BMTH200
BMTH600
BMTH800

FIGURE 10: A METHOD EQUATING FALL 1979 SAMPLES



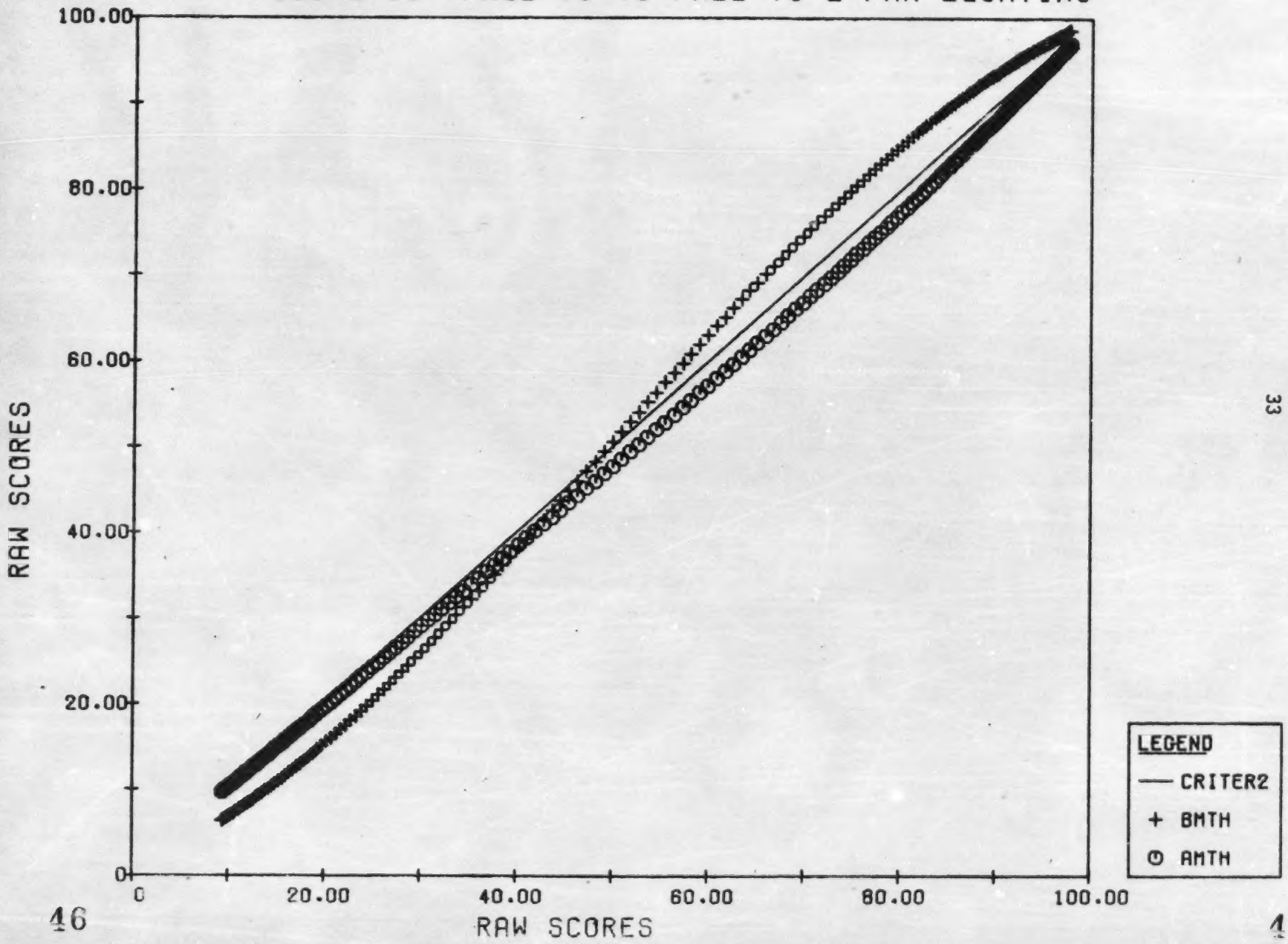
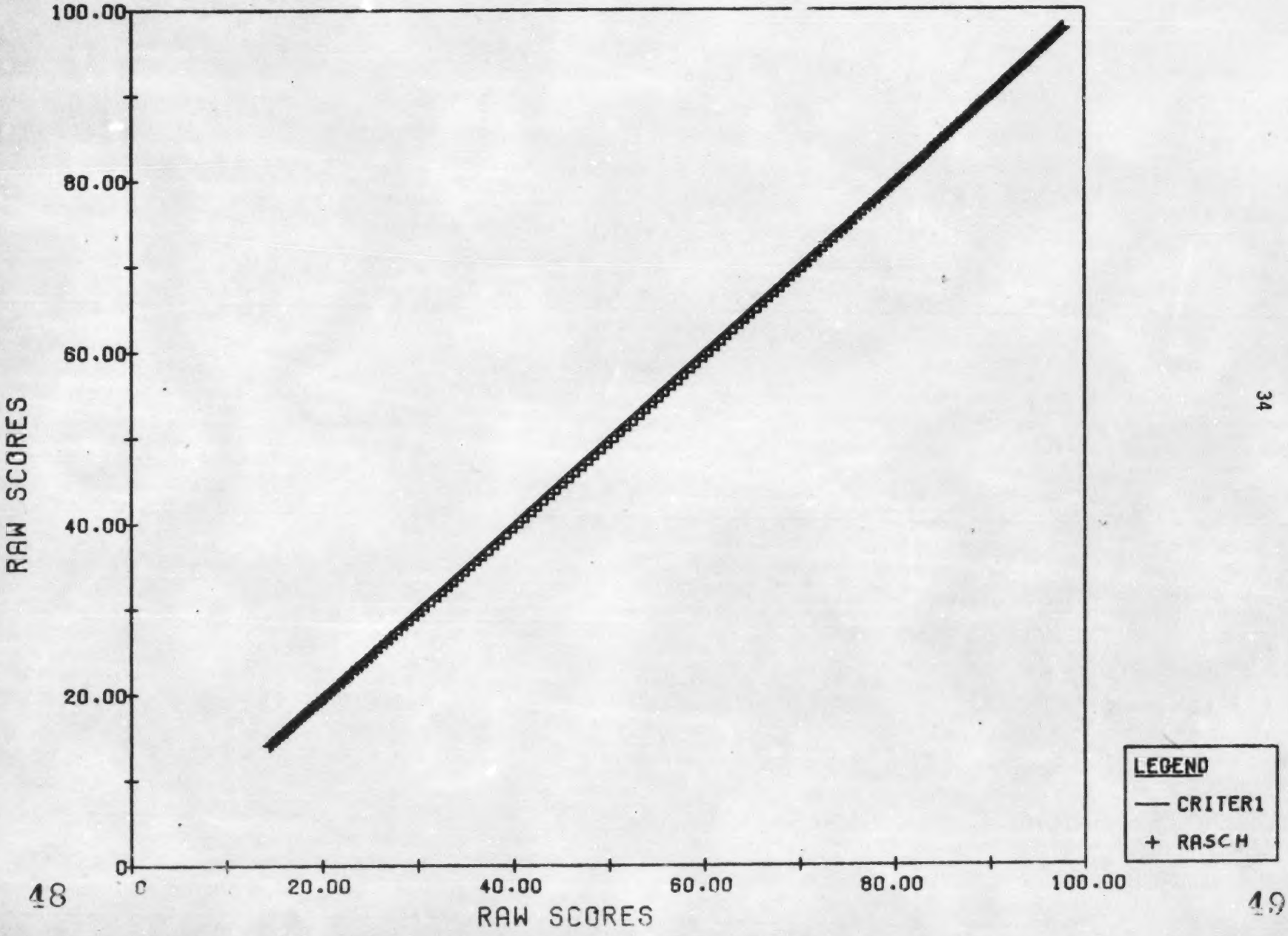


FIGURE 11: FALL 79 TO FALL 79 2 PAR EQUATING
100.00
80.00
60.00
40.00
20.00
0
RAW SCORES
0
20.00
40.00
60.00
80.00
100.00
RAW SCORES
LEGEND
CRITER2
BMTH
AMTH

FIGURE 12: FALL 79 TO FALL 79 RASCH EQUATING



virtually the same, with a slight but consistent difference. The difference is less than .8 of a raw score unit at all points along the scale.

Discussion

Somewhat surprisingly, the two-parameter model calibration method developed by the author provides equating results that are more consistent and valid than those resulting from the regression of item difficulties. It was originally felt that since the item difficulties should be better estimated than the item discriminations, an estimate based on item difficulties would be better. However, in retrospect the results do not seem as surprising. Although the multiplicative constant in the calibrating equation is based entirely on the item discriminations, the additive constant is based on the multiplicative constant and the average of the item difficulties from each pair of tests. Therefore the alternative method uses all of the information contained in the item statistics for calibration while the item difficulty regression approach uses only the item difficulties. It should be mentioned that although the alternative method uses all of the difficulty and discrimination values, it is only a heuristic estimation procedure. Other calibration methods should be devised and investigated.

Another surprising result was the clear superiority of the Rasch model for equating. One would expect the two-parameter model with less restrictions than the Rasch to perform better. The only reasonable explanation seems to be that there is some problem with the estimation of the parameters associated

with the two-parameter model. Perhaps there was not enough data to accurately estimate the additional parameters of the two-parameter model. Examinee sample sizes of between 200 and 1082 may not be large enough for the two-parameter model. Another possible reason is that estimation problems still exist for models with more than one item parameter. Some would say that these estimation problems are intrinsic to logistic models other than the Rasch (e.g., see Wright, 1977, p. 102-104). It is also possible that the Rasch calibration may have been better than the two-parameter model calibration.

Although a test of model validity was provided by equating the Fall 1979 examination to itself, it should be noted that equating a test to itself is somewhat artificial. All items, not just those in the calibrating sections are actually the same across examinee samples. This removes the error that might result from different items in the two tests. Also the data were from the same administration and from roughly equivalent examinee groups. It is correct to infer from the results that the Rasch model was superior to the two-parameter model. However, it is not correct to conclude that equating across examinations, academic terms and students differing in ability would be as good for either model.

Guidelines

The general strategy for investigating and comparing models outlined at the beginning of this paper and in Douglass (1979) worked quite well for comparing the Rasch, two-parameter and three-parameter logistic models. I recommend that this approach be used. However model utility should be

investigated for each practical problem of interest. For example, although equating across terms was the emphasis in this study, further studies are planned for investigating equating of mid-term examinations across instructors and sections.

Some specific guidelines based on our experience are:

1. When considering ICC models include the Rasch model. The computer costs of the Rasch are considerably less than those of the two- and three-parameter models. It is also much easier to work with the simpler Rasch model.
2. Use BICAL to obtain Rasch model estimates. The newer versions of BICAL have many convenient features such as optional deletion of misfitting persons, standard error estimates for item difficulties, listings of item by serial, difficulty and fit order and various useful scattergrams. The only drawbacks are that BICAL currently costs $300 and must be adapted for any systems other than an IBM computer system.
3. When using LOGIST remove poor items before attempting parameter estimation. It was previously mentioned that the difficulties and discriminations for a few items can take on absurd values. LOGIST restricts the movement of any estimate during a given iteration. This can cause the program to continue running for a very long (and expensive) time. Fortunately such items have very low traditional upper-lower discrimination indices. Removal of these items before estimation should solve the problem.

4. Evaluate the models over some time and over normal variations in the practical testing situation. In this study no practical difference in equating results was found until the second year.
5. Do not place too much emphasis on the assumptions *a priori*. This study would have never begun if the assumption of unidimensionality had been taken too literally. An achievement test over a survey course containing at least ten distinct topics does not appear unidimensional. Nor do four-option multiple choice items appear consistent with the Rasch or two-parameter model with lower asymptotes of zero. Yet if only the three-parameter model had been tested, only negative results would have been obtained. However, one cost of not clearly meeting the assumptions of a model is that results must be investigated in each practical situation and for each desired use of the model.
6. It is reasonable to attempt classroom achievement test equating with the Rasch Model with as few as 200 examinees and about 40 of 100 items in common across examinations.
7. With the type of achievement test used in this study, the two-parameter model is not likely to provide consistent or accurate equating with 200 examinees. Six hundred does seem sufficient.
8. Use the calibrating method presented by the author for the two-parameter model. This results in more consistent and accurate calibration than does the method based on the regression of item difficulties.

9. Generally, with smaller sample sizes try the Rasch model. It may be the only model that will provide consistent results.

Conclusion

It is apparent that item characteristic curve models can be successfully applied to classroom achievement testing systems. Further research is needed in other subject areas with various sample sizes and item types. Much work remains to be done using ICC models for general and special purpose test construction. It appears that model assumptions and sample size considerations may not be the limitations they once seemed for classroom achievement testing.

References

Bejar, I., Weiss, D., & Kingsbury, G. Calibration of an item pool for the adaptive measurement of achievement (Research Report No. 77-5). Minneapolis, Minn.: University of Minnesota, Psychometric Methods Program, Department of Psychology, 1977.

Douglass, J. B. A process for testing a mathematical model for the solution of a practical problem: applications to test equating. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 1979. (ERIC Document Reproduction Service No. ED 175 942)

Eisenberg, E. M., & Book, C. Applying latent trait theory to a course examination system: administration, maintenance, and training. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts, April 1980.

Hambleton, R. K., & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement 1977, 14, 75-96.

Hambleton, R. K., Swaminathon, H., Cook, L., Eignor, D., & Gifford, J. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (ETS RB 75-33). Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Rasch, G. Probabilitistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedogogiske Institut, 1960.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Research Memorandum 76-6). Princeton, N. J.: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., Mead, R., & Bell, S. BICAL: calibrating items with the Rasch model. (Research Memorandum No. 23B). Chicago: University of Chicago, Statistical Laboratory, Department of Education, 1979.

Wright, B. D., & Stone, M. Best test design: Rasch measurement. Chicago, Ill.: MESA, 1979.